THE EFFECT OF LANGUAGE FACTORS FOR ROBUST SPEAKER RECOGNITION

Liang Lu[1], Yuan Dong[1, 2], Xianyu Zhao[2], Jiqing Liu[1], Haila Wang[2]

[1]Beijing University of Posts and Telecommunications, Beijing, 100876, China
[2]France Telecom Research & Development Center, Beijing, 100083, China

luliang07@gmail.com xianyu.zhao@orange-ftgroup.com

ABSTRACT

From the results of the NIST speaker recognition evaluation in resent years, speaker recognition systems which are mainly developed based on English training data suffer the language gap problem, namely, the performance of non-English trails is much worse than that of English trails. This problem is addressed in this paper. Based on the conventional joint factor analysis model, we enrolled in the language factors which are mean to capture the language character of each testing and training speech utterance, and compensation was carried out by removing the language factors in order to shrink the difference between languages. Experiments on 2006 NIST SRE data show that, the language factor compensation alone can reduce the gap between the performance of English and non-English trails, and the score level combination with eigenchannels can further improve the performance of non-English trails, e.g., for female part, we observed about 19% relatively reduction in EER, when compared with eigenchannels session variability compensation alone.

Index Terms— Speaker recognition, Joint Factor Analysis, Eigenchannels, Language Factor Compensation

1. INTRODUCTION

In recent years, automatic speaker recognition has achieved great progress, partly driven by NIST speaker recognition evaluation, and also some new techniques which have been demonstrated effective on this task. Those techniques cover most important issues in speaker recognition, such as speaker modeling [1], session variability compensation [2, 3], system fusion [4] and discriminative model training [2, 5], etc. One of the techniques which have successfully applied in the task is joint factor analysis outlined by Kenny, *et al.* [6], which combines relevance MAP, eigenvoices adaptation and eigenchannels session variability compensation into a unified framework.

Although these improvements are remarkable, however, from the results of NIST speaker recognition evaluation in recent years, the language gap problem still troubles most speaker recognition systems, namely, the performance of non-English trails are much worse than that of English trails [7]. This threatens the robustness of recognition systems. The causes are manifold, such as different characteristics between languages themselves. But the main reason may lies in the fact that, the whole system is built up mainly based on English development data, and hence for non-English trails, there is a language mismatch which causes the performance degradation.

In most cases, collecting large amount data from various languages is a very difficult task, hence, this problem can not be solved easily by improving the development data's coverage of languages. Based on the framework of joint factor analysis, in this paper, we investigated the probability of enrolling in a language factor in JFA model in order to capture the language character of training and testing utterances. Just like eigenvoices and eigenchannels, in this approach, a low dimensional subspace was estimated from a multi-language dataset, which describe the main directions of language variation, and the language factors compensation is carried out by removing from the utterances their attribute in this subspace.

The reminder of the paper is organized as follows. Section 2 generally describes joint factor analysis model for speaker verification, and in section 3, we discuss the language factors in JFA in detail. Experimental results are presented in section 4 and section 5 summarizes the paper as a conclusion.

2. JOINT FACTOR ANALYSIS MODEL

The factor analysis techniques outlined by Kenny, *et al.* [6] can be seen as an extension for conventional GMM-UBM approach, which joints the eigenvoices speaker modeling, eigenchannels variability compensation and relevance MAP adaptation into a unified framework. As such, a GMM supervector is decomposed into speaker- and session-dependent parts, which are treated as Gaussian distributed, respectively. The motivation behind this is to explicitly model and separate the speaker and session contributions.

If we let F be the dimension of the acoustic feature vectors used, and C as the total number of Gaussian mixture components, then a GMM mean supervector is $CF \times 1$ dimensional, formed by concatenating the component means together, which can be expressed as $\mu = \left[\mu_1^* \cdots \mu_c^*\right]$, where μ_c is the c^{th} component mean.

For joint factor analysis, a speaker- and channel-dependent supervector M can be decomposed into a sum of two supervectors, namely, a speaker supervector s and a channel supervector c :

$$M = s + c \tag{1}$$

where s and c are statistically independent and normally distributed. The speaker-dependent GMM mean supervector can be represented by

$$s = m + Vy + Dz \tag{2}$$

In this model, m is a $CF \times 1$ supervector which is speaker-independent, representing the center of model space; V is a rectangular matrix of low rank and y is a normally distributed random vector; D is a $CF \times CF$ diagonal matrix and z is a normally distributed CF -dimensional random vector. V and y are commonly referred as eigenvoices and speaker factors,

respectively. It is assumed that the majority of speaker variation is contained within the low-rank speaker subspace defined by VV^*, while the role Dz is mean to model the residual variability that is not captured by the speaker subspace.

A similar expression is used to describe the channel-dependent supervectors:

$$c = Ux \tag{3}$$

where U is a rectangular matrix of low rank which represents the main directions of session variation. The vector x is an estimate of the session conditions with the session subspace, and follows a standard normal distribution. Similar as the speaker subspace, U is often referred as eigenchannels and x as channel factors.

3. LANGUAGE FACTORS IN JFA MODEL

In joint factor analysis model, a speaker's GMM mean supervector is separated into speaker- and session-dependent parts with the hope that each contribution can be estimated and modeled more explicitly. However, if the speaker subspace VV^* and session subspace UU^* are estimated mainly from English data, then the effectiveness of JFA on non-English trail testing is doubtful. In this section, we consider the case of adding a language factor in JFA model which mean to capture and compensate the language character of an utterance in order to handle the language mismatch problem.

3.1. Extend JFA model with language factors

By involving a language factor, we can assume that the speaker subspace VV^* and session subspace UU^* are language-independent which can be estimated by a pure language data, and all the differences between languages can be captured by the language factors. In this case, a speaker's GMM mean supervector can be split as follows:

$$M = m' + Bg + Vy + Dz + Ux \tag{4}$$

where m' is a speaker and language-independent supervector, B is low-rank rectangular transformation matrix, which captures the main directions of language variation. The vector g is an estimation of language condition and is also standard normal distributed. For consistency, we refers the BB^* as language subspace and g as the language factors.

3.2. Language subspace estimation

As for the estimation of language subspace, the multi-language data should first remove their speaker and session attribute. This can adopt the procedure of decoupled estimation of D and V in [8]. However, if the amount of each language data is sufficient large, then the speaker factors can be averaged out. In this case, we can assume that $V = 0, U = 0$ and $D = 0$, then the procedures of training language subspace could be simplified as the eigenvoices modeling addressed in [9] and this approach was adopted in this paper. Here, we give a brief description: given the current estimation of language subspace BB^* (B can be randomly initialized), for each language l, the posterior distribution of its language factor $g(l)$ conditioned on the acoustic observations $\chi(l)$ is Gaussian with mean $\zeta(l)^{-1} B^* \Sigma^{-1} \tilde{F}(l)$ and variance $\zeta(l)^{-1}$, where $\zeta(l) = I + B^* \Sigma^{-1} N(l) B$, and $N(l), \tilde{F}(l)$ is centralized zero- and first order Baum-Welch statistics, respectively. The re-estimation of language subspace basis B involves the EM iteration which is mean to maximize the following likelihood function:

$$\prod_l \max_g P_{HMM}\left(\chi(l) \middle| m' + Bg, \Sigma\right)$$

where l ranges over all the languages. It should be noted that, for small number of languages, the amount of each should be balanced in order to avoid that those languages with large amount data will dominate the estimation of language subspace.

3.3. Language factor compensation (LFC)

After the estimation of language subspace, the next procedure is to compensate the language factors in joint factor analysis model. Generally, there are two approaches for this task:

1) The first way is to incorporate the language factors with speaker factors for speaker modeling, then the speaker dependent supervector would be:

$$s' = m' + Bg + Vy + Dz \tag{5}$$

This approach can model both the language and speaker characters for a speaker simultaneously; however, it may enroll in the intra-speaker variability for bi-linguistic speakers, who frequently appear in Mixer corpus [7].

2) Another strategy is to maintain the speaker dependent supervector as (2) unchanged, while combine the language factors with channel factors as nuisance attribute which is removed from the speaker's supervectors:

$$c' = Bg + Ux \tag{6}$$

The motivation of this approach lies in that, the mismatch between English data developed recognition systems and non-English utterances will be alleviated by removing the language characters. In this paper, we adopted the second approach for language factor compensation.

4. EXPERIMENTAL RESULTS

In this section, we report speaker verification experiments on the 2006 NIST SRE corpus with language factors compensation (LFC). In this paper, we focus the discussion on the effect of language factors and the comparison with eigenchannels. To make things clear, we did not consider the role of eigenvoices, i.e. $V = 0$, in all the experiments. Section 4.1 presents some general experiment setup information about the task, database and features. The results of these experiments are discussed in section 4.2.

4.1. Protocol

Speaker verification experiments were conducted on the 2006 NIST SRE corpus [10]. We focused on the core condition task, which involves 3,612 true trials and 47,836 false trials. The number of English trails is 24013 while that of non-English trails is 27435. Enrollment and testing utterance contain about 2 minutes of pure speech after some voice activity detection.

	English	Farsi	French	German	Hindi	Japanese	Korean	Mandarin	Thai	Total size
OGI	2.60	1.29	1.36	1.33	0.24	0.98	1.01	1.16	-	10.58
Mixer	3.06	0.95	-	-	6.26	3.97	1.95	1.87	3.38	21.44
Total size	5.66	2.24	1.36	1.33	6.50	4.95	2.96	3.03	3.38	32.02
	Tamil	Spanish	Vietnam	Bengali	Russian	Italian	Arabic	Tagalog	Cantonese	Total size
OGI	1.33	1.56	0.95	-	-	-	-	-	-	3.84
Mixer	-	1.72	6.95	1.32	4.65	1.29	3.21	1.31	5.48	25.93
Total size	1.33	3.28	7.90	1.32	4.65	1.29	3.21	1.31	5.48	29.77

Table 1. The category of languages used in the experiments and its corresponding amount of speech (after voice activity detection and measured in the unit of hour), in which, the total amount of OGI multi-language data is 14.42 hours, and 47.37 hours for Mixer data.

4.1.1. Database

The universal background model (UBM) is trained on Switchboard and Mixer corpus. The majority of the data is spoken in English, while only a small portion of Mixer corpus data is spoken by several other languages. NIST SRE 2004 and 2005 telephone data is used to model the eigenchannels, which contains thousands of utterances coming from 829 speakers. In order to model the language subspace, we selected the multi-language data from two corpuses: one is the Oregon Graduate Institute (OGI) multi-language corpus [11], which covers only 11 languages and much of the languages have small amount of data. To enlarge the amount as well as the diversity of the multi-language dataset, we also selected some data from 2004 and 2008 NIST SRE data, which come from Mixer corpus. The number of languages we used is 18 including English, and the total amount is about 62 hours. The detailed information about this corpus is presented in table 1.

4.1.2. System configuration

For the features in the experiment, 12 MFCC coefficients plus C0 are computed and cepstral mean subtraction (CMS) and feature warping over 300 frames are applied. RASTA filtering of the features follows. First, second and third order derivatives computed over 5 frames are appended to each feature vector, which results in dimensionality 52. HLDA is used to reduce the feature dimension from 52 to 39. A gender-independent UBM is used and its number of Gaussian components is 2048. The rank of eigenchannels basis U is 100 and that of language subspace B is 12. In our experiments, B is trained by EM iterations while U is simply obtained by kernel principle component analysis.

4.2. Results

Figure 1 and 2 show the DET curves of male and female part speaker verification results on the 1conv4w-1conv4w task of 2006 NIST SRE. The baseline system uses the relevance MAP adaptation only, namely, $D^2 = \frac{1}{r}\Sigma$ and no session or language compensation was adopted. The relevance factor r was set to be 16. The comparison experiments use language factor compensation (LFC) in both training data testing phases [12]. In training phase, the language factors of training utterances were removed from the models, and in testing phase, compensation was performed in the model level, namely:

$$llr(X_{utt}, M_{tar}) = \frac{1}{T}\log\left(\frac{p(X_{utt}|M_{tar} + Bg_h)}{p(X_{utt}|M_{UBM} + Bg_h)}\right) \quad (7)$$

where g_h denotes the language factor of test utterance X_{utt}.

We can see from the results that, on both male and female trails test, the language factor compensation is more effective for non-English trails when compared with English trails, and the gap between the two is narrowed accordingly. This may due to the fact that the language factor compensation (LFC) reduced the mismatch between English data trained UBM and non-English utterances.

For the further discussion, we performed the second set of experiments on female part to compare the performance of language factor compensation (LFC) with eigenchannels as well as the combination of the two. The compensation in all of those experiments were also carried out in both training and testing phases, and two kinds of combination approach were compared, namely, model level combination as equation (6) shows, and score level fusion as follows:

$$s(X_{utt}) = \alpha \cdot llr_{ec}(X_{utt}) + (1-\alpha) llr_{lfc}(X_{utt}) \quad (8)$$

where $\alpha \subset [0,\ 1]$ is the weight parameter.

The results are presented in table 2 in both EER and DCF, which shows that, eigenchannels session variability compensation improves the performance of English trails significantly, in both EER and DCF. For non-English trails, however, the improvement is not so remarkable after eigenchannels. The reason may be just as it is discussed in section 3, that the eigenchannels which were trained mainly only on English data can not well describe the property of non-English data. Language factor compensation, on the other hand, can outperform eigenchannels on non-English trails

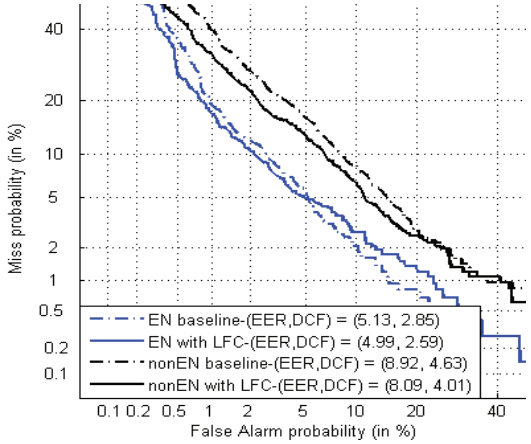

Fig. 1. DET curves before and after the language factor compensation of male part trails, English and non-English trails, respectively.

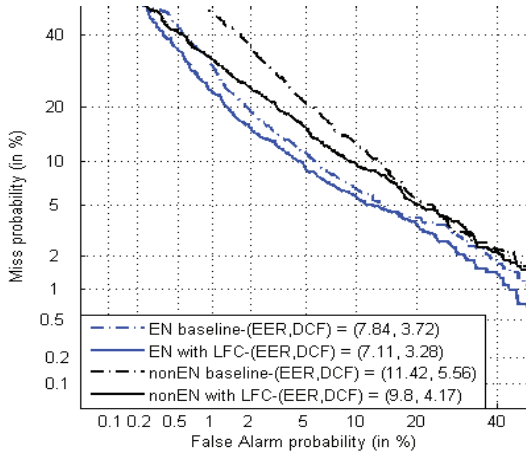

Fig. 2. DET curves before and after the language factor compensation of female part trails, English and non-English trails, respectively.

by only a 12 dimensional subspace trained on a small dataset. Hence, a straightforward idea is to combine the two approaches.

From our results, however, the model level combination of the two did not achieve promising result in both English and non-English trails, whereas the score level fusion obtained significant improvement for non-English trails. The result of score level fusion indicates that the two approaches can be complementary. However, as for the failure of model level combination, the major reason may lies in the fact that the dimension of language subspace is too small and it's noisy when compared with that of session subspace (12 dimensional language subspace roughly trained by 18 different languages vs. 100 dimensional session subspace estimated by thousands of utterances). Additionally, the two subspaces estimated independently can not be guaranteed to cooperate well by pooling the basis together. When using maximum a posterior criteria for model training, it may be able to find a point with higher posteriori likelihood when moving in session subspace while do not consider the role of language subspace. Our future work will further investigate this question and the cooperation of language factor compensation with eigenvoices will also be addressed.

5. CONCLUSIONS

This paper addressed the language gap problem in speaker recognition systems, namely, the systems which are trained mainly on English corpus data performs much worse on non-English trails. We adopted the approach of enrolling in a language factor in joint factor analysis models, whose role is to capture the language characters of training and testing utterances. Language factor compensation was carried out by removing the language attribute in both training and testing phases in order to reduce the mismatch between utterances and system. Experiments on 2006 NIST SRE data showed that, the language factor compensation itself can significantly improve the performance of non-English trails and meanwhile, narrow the gap of performance between English and non-English trails. When combined with eigenchannels, score level fusion achieved 19% relative improvement on the non-English trails of female part, when compared with eigencahnnels alone. Future works will focus on the model level combination of language factor compensation and eigenchannels as well as eigenvoices.

Systems	English trails		non-English trails	
	EER	DCF	EER	DCF
Baseline	7.84%	.372	11.42%	.566
LFC only	7.11%	.328	9.8%	.417
eigenchannels only	5.03%	.223	11.19%	.412
Combination in model level	5.13%	.226	11.19%	.408
Combination in score level	**5.13%**	**.218**	**9.04%**	**.374**

Table 2. Comparison on female part of language factor compensation and session variability compensation as well as the different combination approaches of the two.

6. REFERENCES

[1] W. M. Campbell, D. E. Sturim, D. A. Reynolds and A. Solomonoff, "SVM-based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. ICASSP*, 2006, pp. 97-100.

[2] A. Solomonoff, W. M. Campbell and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proc. ICASSP*, 2005, pp. 629-632.

[3] X. Zhao, Y. Dong, H. Yang, J. Zhao, L. Lu and H. Wang, "Nonlinear kernel nuisance attribute projection for speaker verification," in *proc. ICASSP*, 2008, pp. 4125-4128.

[4] N. Brummer, L. Burget and J. Cernocky *et al.* "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Trans. Speech Audio Processing,* vol. 15, no. 7, pp. 2072-2084, Sept. 2007.

[5] L. Lu, Y. Dong, X. Zhao, J. Zhao, C. Dong and H. Wang, "Analysis of subspace within-class covariance normalization for SVM based speaker verification," in *proc. Interspeech,* 2008.

[6] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms, Tech. Report GRIM-06/08-13," 2005.

[7] M. A. Prezybocki, A. F. Martin, and A. N. Le, "NIST speaker recognition evaluations utilizing the mixer corpora-2004, 2005, 2006," *IEEE Trans. Speech Audio Processing,* Sept. 2007.

[8] P. Kenny, P. Ouellet, N. Dehark, V. Gupta, and P. Dumouchel, "A Study of Inter-Speaker Variability in Speaker Verification," *IEEE Trans. Speech Audio Processing,* July 2008.

[9] P.Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Processing,* vol. 13, no. 3, pp. 345-359, May 2005.

[10] "The NIST 2006 speaker recognition evaluation plan," http://www.nist.gov/speech/tests/spk/spk/2006.

[11] Y. K. Muthusamy, R. A. Cole and B. T. Oshika, "The OGI Multi-language Telephone Speech Corpus," in *Proc. ICSLP,* Banff, Alberta, Canada, October 1992.

[12] D. Matrouf, N. Scheffer, B. Fauve and J.-F. Bonastre, "A straight and efficient implementation of the factor analysis model for speaker verification," *in Proc. Interspeech*, 2007, pp. 1242-1245.